NULEX: An Open-License Broad Coverage Lexicon

Clifton J. McFate
Northwestern University
Evanston, IL. USA.
c-mcfate@northwestern.edu

Kenneth D. Forbus
Northwestern University
Evanston, IL. USA
forbus@northwestern.edu

Abstract

Broad coverage lexicons for the English language have traditionally been handmade. This approach, while accurate, requires too much human labor. Furthermore, resources contain gaps in coverage, contain specific types of information, or are incompatible with other resources. We believe that the state of open-license technology is such that a comprehensive syntactic lexicon can be automatically compiled. This paper describes the creation of such a lexicon, NU-LEX, an open-license feature-based lexicon for general purpose parsing that combines WordNet, VerbNet, and Wiktionary and contains over 100,000 words. NU-LEX was integrated into a bottom up chart parser. We ran the parser through three sets of sentences, 50 sentences total, from the Simple English Wikipedia and compared its performance to the same parser using Comlex. Both parsers performed almost equally with NU-LEX finding all lex-items for 50% of the sentences and Comlex succeeding for 52%. Furthermore, NULEX's shortcomings primarily fell into two categories, suggesting future research directions.

1 Introduction

While there are many types of parsers available, all of them rely on a lexicon of words, whether syntactic like Comlex, enriched with semantics like WordNet, or derived from tagged corpora like the Penn Treebank (Macleod *et al*, 1994; Fellbaum, 1998; Marcus et al, 1993). However, many of these resources have gaps that the others can fill in. WordNet, for example, only contains open-class words, and it lacks the extensive subcategorization frame and agreement information present in Comlex (Miller *et al*, 1993; Macleod *et al*, 1994). Comlex, while syntactically deep, doesn't have tagged usage data or semantic groupings (Macleod *et al*, 1994). Furthermore, many of these resources do not map to one another or have restricted licenses.

The goal of our research was to create a syntactic lexicon, like Comlex, that unified multiple existing open-source resources including Felbaum's (1998) WordNet, Kipper *et al*'s (2000) VerbNet, and Wiktionary. Furthermore, we wanted it to have direct links to frame semantic representations via the open-license OpenCyc knowledge base.

The result was NU-LEX a lexicon of over 100,000 words that has the coverage of WordNet, is enriched with tense information from automatically screen-scrapping Wiktionary[1], and contains VerbNet subcategorization frames. This lexicon was incorporated into a bottom-up chart parser, EANLU, that connects the words to Cyc representations (Tomai & Forbus 2009). Each entry is represented by Cyc assertions and contains syntactic information as a set of features consistent with previous feature systems (Allen 1995; Macleod *et al*, 1994).

[1] http://www.wiktionary.org/

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 363–367,
Portland, Oregon, June 19-24, 2011.

2 Previous Work

Comlex is handmade and contains 38,000 lemmas. It represents words in feature value lists that contain lexical data such as part of speech, agreement information, and syntactic frame participation (Macleod *et al*, 1994). Furthermore, Comlex has extensive mappings to, and uses representations compatible with, multiple lexical resources (Macleod *et al*, 1994).

Attempts to automatically create syntactic lexical resources from tagged corpora have also been successful. The Penn Treebank is one such resource (Marcus et al, 1993). These resources have been successfully incorporated into statistical parsers such as the Apple Pie parser (Sekine & Grishman, 1995). Unfortunately, they still require extensive labor to do the annotations.

NU-LEX is different in that it is automatically compiled without relying on a hand-annotated corpus. Instead, it combines crowd-sourced data, Wiktionary, with existing lexical resources.

This research was possible because of the existing lexical resources WordNet and VerbNet. WordNet is a virtual thesaurus that groups words together by semantic similarity into synsets representing a lexical concept (Felbaum, 1998). VerbNet is an extension of Levin's (1993) verb class research. It represents verb meaning in a class hierarchy where each verb in a class has similar semantic meanings and identical syntactic usages (Kipper *et al*, 2000). Since its creation it has been expanded to include classes not in Levin's original research (Kipper *et al*, 2006). These two resources have already been mapped, which facilitated applying subcategorization frames to WordNet verbs.

Furthermore, WordNet has existing links to OpenCyc. OpenCyc is an open-source version of the ResearchCyc knowledge base that contains hierarchical definitional information but is missing much of the lower level instantiated facts and linguistic knowledge of ResearchCyc (Matuszek *et al*, 2006). Previous research by McFate (2010) used these links and VerbNet hierarchies to create verb semantic frames which are used in EANLU, the parser NU-LEX was tested on.

3 Creating NU-LEX

The NU-LEX describes words as CycL assertions. Each form of a word has its own entry. For the purposes of integration into a parser that already uses Comlex, the formatting was kept similar. Because the lexification process is automatic, formatting changes are easy to implement.

3.1 Nouns

Noun lemmas were initially taken from Fellbaum's (1998) WordNet index. Each Lemma was then queried in Wiktionary to retrieve its plural form resulting in a triple of *word*, *POS*, and *plural form*:

```
(boat Noun (("plural" "boats")))
```

This was used to create a definition for each form. Each definition contains a list of WordNet *synsets* from the original word, the *orthographic word form* which was assumed to be the same as the word, *countability* taken from Wiktionary when available, the *root* which was the base form of the word, and the *agreement* which was either singular or plural.

```
(definitionInDictionary WordNet "Boat"
  (boat (noun
  (synset ("boat%1:06:01:"
          "boat%1:06:00::"))
  (orth "boat")
  (countable +)
  (root boat) (agr 3s))))
```

3.2 Verbs

Like Nouns, verb base lemmas were taken from the WordNet index. Similarly, each verb was queried in Wiktionary to retrieve its tense forms resulting in a list similar to that for nouns:

```
(give Verb ((
("third-person singular simple present"
"gives")
("present participle" "giving")
("simple past" "gave")
("past participle" "given"))))
```

These lists in turn were used to create the *word*, *form*, and *agreement* information for a verb entry. The *subcategorization* frames were taken directly from VerbNet. *Root* and *Orthographical form* were again kept the same.

```
(definitionInDictionary WordNet "Give"
  (give (verb
    (synset ("give%2:41:10::…
            …"give%2:34:00::"))
    (orth "give")
    (vform pres)
    (subcat (? S np-v-np-np-pp.asset
      np-v-np-pp.recipient-pp.asset
      np-v-np-pp.asset
      np-v-pp.recipient
      np-v-np
      np-v-np-dative-np
```

```
    np-v-np-pp.recipient))
  (root give)
  (agr (? a 1s 2s 1p 2p 3p)))))
```

3.3 Adjectives and Adverbs

Adjectives and adverbs were simply taken from WordNet. No information from Wiktionary was added for this version of NU-LEX, so it does not include comparative or superlative forms. This will be added in future iterations by using Wiktionary. The lack of comparatives and superlatives caused no errors. Each definition contains the *Word*, *POS*, and *Synset list*:

```
(definitionInDictionary WordNet "Funny"
  (funny (adjective
    (root funny)
    (orth "funny")
    (synset ("funny%4:02:01::"
             "funny%4:02:00::")))))
```

3.4 Manual Additions

WordNet only contains open-class words: Nouns, Adjectives, Adverbs, and Verbs (Miller *et al*, 1993). Thus determiners, subordinating conjunctions, coordinating conjunctions, and pronouns all had to be hand created.

Likewise, Be-verbs had to be manually added as the Wiktionary page proved too difficult to parse. These were the only categories added.

Notably, proper names and cardinal numbers are missing from NU-LEX. Numbers are represented as nouns, but not as cardinals or ordinals. These categories were not explicit in WordNet (Miller *et al*, 1993).

4 Experiment Setup

The sample sentences consisted of 50 samples from the Simple English Wikipedia[2] articles on the heart, lungs, and George Washington. The heart set consisted of the first 25 sentences of the article, not counting parentheticals. The lungs set consisted of the first 13 sentences of the article. The George Washington set consisted of the first 12 sentences of that article. These sets corresponded to the first section or first two sections of each article. There were 239 unique words in the whole set out of 599 words total.

Each set was parsed by the EANLU parser. EANLU is a bottom-up chart parser that uses compositional semantics to translate natural language into Cyc predicate calculus representations (Tomai & Forbus 2009). It is based on a Allen's (1995) parser. It runs on top of the FIRE reasoning engine which it uses to query the Cyc KB (Forbus *et al*, 2010).

Each sentence was evaluated as correct based on whether or not it returned the proper word forms. Since we are not evaluating EANLU's grammar, we did not formally evaluate the parser's ability to generate a complete parse from the lex-items, but we note informally that parse completeness was generally the same. Failure occurred if any lex-item was not retrieved or if the parser was unable to parse the sentence due to system memory constraints.

5 Results

Can NU-LEX perform comparably to existing syntactic resources despite being automatically compiled from multiple resources? Does its increased coverage significantly improve parsing? How accurate is this lexicon?

In particular we wanted to uncover words that disappeared or were represented incorrectly as a result of the screen-scraping process.

Overall, across all 50 samples NU-LEX and Comlex performed similarly. NULEX got 25 out of 50 (50%) correct and Comlex got 26 out of 50 (52%) of the sentences correct. The two systems made many of the same errors, and a primary source of errors was the lack of proper nouns in either resource. Proper nouns caused seven sentences to fail in both parsers or 29% of total errors.

Of the NU-LEX failures not caused by proper nouns, five of them (20%) were caused by lacking cardinal numbers. The rest were due to missing lex-items across several categories. Comlex primarily failed due to missing medical terminology in the lungs and heart test set.

Out of the total 239 unique words, NULEX failed on 11 unique words not counting proper nouns or cardinal numbers. One additional failure was due to the missing pronoun "*themselves"* which was retroactively added to the hand created pronoun section. This a failure rate of 4.6%. Comlex failed on 6 unique words, not counting proper nouns, giving it a failure rate of 2.5%.

5.1 The Heart

For the heart set 25 sentences were run through the parser. Using NU-LEX, the system correctly identified the lex-items for 17 out of 25 sentences (68%). Of the sentences it did not get correct, five were incorrect only because of the

[2] http://simple.wikipedia.org/wiki/Main_Page

lack of cardinal number representation. One failed because of system memory constraints.

Using Comlex, the parser correctly identified all lex-items for 16 out of 25 sentences (64%). The sentences it got wrong all failed because of missing medical terms. In particular, *atrium* and *vena cava* caused lexical errors.

5.2 The Lungs

For the lung set 13 sentences were run through the parser. Using NU-LEX the system correctly identified all lex-items for 6 out of 13 sentences (46%). Two errors were caused by the lack of cardinal number representation and one sentence failed due to memory constraints. One sentence failed because of the medical specific term *para-bronchi*.

Four additional errors were due to a malformed verb definitions and missing lexitems lost during screen scraping.

Using Comlex the parser correctly identified all lex-items for 7 out of 13 sentences (53%). Five failures were caused by missing lex-items, namely medical terminology like *alveoli* and *parabronchi*. One sentence failed due to system memory constraints.

5.3 George Washington

For the George Washington set 12 sentences were run through the parser. This was a set that we expected to cause problems for NU-LEX and Comlex because of the lack of proper noun representation. NU-LEX got only 2 out of 12 correct and seven of these errors were caused by proper nouns such as *George Washington*.

Comlex did not perform much better, getting 3 out of 12 (25%) correct. All but one of the Comlex errors was caused by missing proper nouns.

6 Discussion

NU-LEX is unique in that it is a syntactic lexicon automatically compiled from several open-source resources and a crowd-sourced website. Like these resources it too is open-license. We've demonstrated that its performance is on par with existing state of the art resources like Comlex. By virtue of being automatic, NU-LEX can be easily updated or reformatted. Because it scrapes Wiktionary for tense information, NU-LEX can constantly evolve to include new forms or corrections. As its coverage (over 100,000 words) is derived from Fellbaum's (1998) WordNet, it is also significantly larger than existing similar syntactic resources.

NU-LEX's first trial demonstrated that it was suitable for general purpose parsing. However, much work remains to be done. The majority of errors in the experiments were caused by either missing numbers or missing proper nouns. Cardinal numbers could be easily added to improve performance. Furthermore, solutions to missing numbers could be created on the grammar side of the process.

Missing proper nouns represent both a gap and an opportunity. One approach in the future could be to manually add important people or places as needed. Because the lexicon is Cyc compliant, other options could include querying the Cyc KB for people and then explicitly representing the examples as definitions. This method has already proven successful for EANLU using ResearchCyc, and could transfer well to OpenCyc. Screen-scraping Wiktionary could also yield proper nouns.

With proper noun and number coverage, total failures would have been reduced by 48%. Thus, simple automated additions in the future can greatly enhance performance.

Errors caused by missing or malformed definitions were not abundant, showing up in only 12 of the 50 parses and under half of the total errors. The total error rate for words was only 4.6%. We believe that improvements to the screen-scrapping program or changes in Wiktionary could lead to improvements in the future.

Because it is CycL compliant the entire lexicon can be formally represented in the Cyc knowledge base (Matuszek *et al*, 2006). This supports efficient reasoning and allows systems that use NU-LEX to easily make use of the Cyc KB. It is easily adaptable in LISP or Cyc based applications. When partnered with the EANLU parser and McFate's (2010) OpenCyc verb frames, the result is a semantic parser that uses completely open-license resources.

It is our hope that NU-LEX will provide a powerful tool for the natural language community both on its own and combined with existing resources. In turn, we hope that it becomes better through use in future iterations.

References

Allen, James. 1995. *Natural Language Understanding: 2nd edition.* Benjamin/Cummings Publishing Company, Inc. Redwood City, CA.

Fellbaum, Christiane. Ed. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Forbus, K., Hinrichs, T., de kleer, J., and Usher, J. 2010.FIRE: Infrastructure for Experience-based Systems with Common Sense. *AAAI Fall Symposium on Commonsense Knowledge.* Menlo Park, CA. AAAI Press.

Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. In *Fifth International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy.

Levin, Beth. 1993. *English Verb Classes and Alternation: A Preliminary Investigation*. The University of Chicago Press, Chicago.

Macleod, Catherine, Ralph Grishman, and Adam Meyers. 1994 Creating a Common Syntactic Dictionary of English. Presented at *SNLR: International Workshop on Sharable Natural Language Resources*, Nara, Japan.

Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics.* 19(2): 313-330.

Matuszek, Cynthia, John Cabral, Michael Witbrock, and John DeOliveira. 2006. An Introduction to the Syntax and Content of Cyc. In *Proceedings of the* 2006 *AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, Stanford, CA.

McFate, Clifton. 2010. Expanding Verb Coverage in Cyc With VerbNet. *In proceedings of the ACL 2010 Student Research Workshop*l. Uppsala, Sweden,

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to WordNet: An On-line Lexical Database. In Fellbaum, Christiane. Ed. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Sekine, Satoshi, and Ralph Grishman. 1995. A Corpus-based Probabilistic Grammar with Only Two Non-terminals. In *Fourth International Workshop on Parsing Technologies*. Prague, Czech Republic.

Tomai, Emmet, and Kenneth Forbus. 2009. EA NLU: Practical Language Understanding for Cognitive Modeling. In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*, Sanibel Island, FL.